Zachodniopomorski Uniwersytet Technologiczny

Streszczenie Pracy Doktorskiej

Modelowanie Call Center z wykorzystaniem niejednorodnych łańcuchów Markowa z czasem ciągłym

Application of inhomogeneous continuous time Markov chains for call center modeling

---

Autor:
Maciej, Rafał Burak

Promotor:
prof. dr hab. inż. Andrzej Brykalski

Wydział Informatyki

11 kwietnia 2018

Spis treści

Rozdział 1

Aktualność problemu

Centra kontaktu z klientem (ang. *call center*) stanowią obecnie integralną część wielu produktów i usług oferowanych zarówno przez przedsiębiorstwa jak również instytucje publiczne. Transformacja sposobu świadczenia usług, z obowiązującej do niedawna formy fizycznego przedstawicielstwa/obecności, w stronę świadczenia usług w sposób zdalny za pomocą call center, rozpoczęła się na świecie w latach osiemdziesiątych a w Polsce na dużą skalę pod koniec lat dziewięćdziesiątych ubiegłego wielu i trwa do dzisiaj. Pozwoliła ona na znaczącą poprawę zarówno standardów obsługi jak również dostępności świadczonych usług. W pierwszej kolejności call center zostały wykorzystane przez przedsiębiorstwa świadczące relatywnie proste usługi na dużą skalę, przede wszystkim w instytucjach finansowych – do masowej obsługi w sposób zdalny klientów detalicznych (bankowość elektroniczna, obsługa kart płatniczych), telekomunikacji – jako podstawowy sposób kontaktu z klientem, biurach podróży, liniach lotniczych i wielu innych. W chwili obecnej trudno sobie wyobrazić możliwość funkcjonowania przedsiębiorstwa w jednej z tych branż, bez wykorzystania usług świadczonych na odległość poprzez call center. Dodatkowym impulsem dla ich rozwoju jest również trend do outsourcingu tego typu usług, dzięki czemu, głównie poprzez wykorzystanie efektu skali i koncentracji know-how, stały się one jeszcze bardziej efektywne, a co za tym idzie możliwe do wykorzystania przez większą liczbę zainteresowanych przedsiębiorstw, bez konieczności drogich inwestycji w infrastrukturę.

Outsourcing usług nie dotyczy jedynie rynku krajowego. Dzięki stosunkowo niskim kosztom płac i dużej bazie wysoko wykwalifikowanej siły roboczej, Polska jest jednym z liderów pozyskiwania inwestycji w obszarze call center lub szerzej BPO (ang. *business process outsourcing*) obsługujących klientów z państw wysoko rozwiniętych. Według raportu firmy ABSL (badanie na zlecenie Polskiej Agencji Informacji i Inwestycji Zagranicznych) z 2014 roku call center / BPO z kapitałem zagranicznym zatrudniają w Polsce ok. 130 tysięcy osób (w tym ok. 60 tys. agentów call center) z perspektywami dalszego dwucyfrowego corocznego wzrostu w najbliższych latach. W całej Unii Europejskiej call center zapewniają 3,2 mln miejsc pracy (>1% zatrudnionych) a w najbardziej rozwiniętych gospodarkach takich jak Stany Zjednoczone, Wielka Brytania czy Holandia do około 3% pracującej populacji (Bain and Taylor, 2002, JLLResearch, 2017).

Dzięki znaczeniu gospodarczemu sektora call center oraz kompleksowości problemów związanych z oferowaniem tego rodzaju usług, stały się one w ostatnich latach tematem wielu publikacji naukowych i długoterminowych projektów badawczych. Przykładowo, opracowania (Aksin, Armony, and Mehrotra, 2007)

i (Gans, Koole, and Mandelbaum, 2003), zawierające przegląd tematyki, powiązanych obszarów badawczych i stosowanych w nich metod, mają odpowiednio ponad 670 i 1420 cytowań, w tym odpowiednio 240 i 308 w latach 2015-2016.

Podstawowym zagadnieniem związanym z zarządzaniem call center, a co za tym idzie z powiązanym obszarem badań, jest planowanie i optymalizacja ilościowa zasobów ludzkich (ang. *capacity management*). Wynika to z jednej strony z faktu, że koszty osobowe stanowią 60-70% całkowitych kosztów call center (Holman, Batt, and Holtgrewe, 2007 lub Gans, Koole, and Mandelbaum, 2003), dodatkowo, w odróżnieniu od tradycyjnych przedsiębiorstw produkcyjnych, praca call center jest zdeterminowana przez stochastyczną i wysoce zmienną w czasie liczbę przychodzących żądań obsługi. Dotyczy to w szczególności call center obsługujących ruch przychodzący (ang. *inbound*), które stanowią większość tego typu organizacji (ok. 78% według Holman, Batt, and Holtgrewe, 2007). Dobór odpowiedniej liczby wykwalifikowanych pracowników w stosunku do przychodzącego – przypadkowego i zmiennego w czasie – ruchu ma kluczowe znaczenie dla możliwości oferowania określonego rodzaju usług na odległość z określoną jakością, w tym również pod postacią tzw. *service level agreement* będących sformalizowaną częścią umów o świadczenie usług. Możliwość lepszego wykorzystania czasu i know-how zatrudnionych pracowników jest również jednym z głównych powodów koncentracji obsługi klienta w przedsiębiorstwach za pomocą call center. Jego planowanie i optymalizacja ma, co za tym idzie, kluczowe znaczenie dla efektywności przedsiębiorstwa.

Znaczenie problemu planowania obsady w systemach obsługi ze stochastycznym i zmiennym w czasie obciążeniem jest odzwierciedlona znaczącą liczbą publikacji poświęconych temu zagadnieniu. Przykładowo, aktualne publikacje – (Defraeye and Nieuwenhuyse, 2015) oraz (Schwarz, Selinka, and Stolletz, 2016) – zawierające przegląd i klasyfikację stosowanych obecnie w tym obszarze metod obejmują odpowiednio 241 i 210 referencji. Przykładem prac prowadzonych w tym obszarze w naszym kraju mogą być publikacje prof. dr hab. inż. Tadeusza Czachórskiego i jego współpracowników (np. Czachórski et al., 2009, Bylina et al., 2009), będące wynikiem badań prowadzonych w kierowanym przez niego Instytucie Informatyki Teoretycznej i Stosowanej PAN.

Ze względu na to, że proces przychodzących żądań obsługi jest wynikiem niezależnych od siebie i występujących z niską częstotliwością decyzji podejmowanych przez populację o dużych rozmiarach, stanowi on klasyczny przykład rozkładu Poissona co powoduje, że systemy obsługi klienta są na ogół modelowane jako Markowskie systemy kolejkowe. Historia ich zastosowań w obszarze badań operacyjnych sięga pionierskich prac A.K. Erlanga z początku XX wieku (Erlang, 1917), których wynikiem był m.in. model Erlang-C opisujący stany ustalone systemu kolejkowego M/M/S. Model ten, jak również w nieco mniejszym zakresie jego rozszerzenie umożliwiające modelowanie rozłączenia się klientów oczekujących w kolejce na połączenie, zaproponowane przez C. Palma (znane jako model Palm/Erlang-A) – również współcześnie dominuje w zastosowaniach praktycznych – głównie ze względu na łatwość implementacji i niską złożoność obliczeniową (Ingolfsson et al., 2007 lub Mandelbaum and Zeltyn, 2007). Co ciekawe, również w przypadku bardziej dokładnych modeli kolejkowych uwzględniających np. inne niż Markowskie, lub zależne od stanu systemu rozkłady procesów obsługi (w tym również opuszczenia systemu bez obsługi) – ich stacjonarne przybliżenia stanowią większość z aktualnie proponowanych w literaturze metod modelowania zmiennych w czasie systemów obsługi, na przy-

kład we wspomnianym aktualnym przeglądzie literatury – (Defraeye and Nieuwenhuyse, 2015), poświęconemu metodom planowania obsady dla zmiennego w czasie stochastycznego obciążenia.

Dokładność modeli zakładających stacjonarny stan systemu może być niewystarczająca w sytuacjach, gdy rzeczywisty stan systemu znacząco odbiega od jego stanu ustalonego (np. przy dużej zmienności procesu przychodzących żądań obsługi i odpowiadających jej znaczących zmian obsady lub przy długich czasach obsługi). Problem adekwatności stacjonarnych przybliżeń modeli kolejkowych używanych do analizy systemów niejednorodnych w czasie był przedmiotem zainteresowania wielu badaczy. Na przykład, w (Green and Kolesar, 1991) oraz (Green, Kolesar, and Soares, 2001) dokonano oceny dokładności metod opartych o przybliżenia stacjonarne kolejkowego modelu M/M/S (Erlang-C) z rozwiązaniem jego reprezentacji jako niejednorodny w czasie łańcuch Markowa czasu ciągłego (ang. *Continuous Time Markow Chain – CTMC*) uwzględniającym stany nieustalone, uzyskanym przez numeryczne rozwiązanie układu równań różniczkowych Chapmana-Kolmogorowa przy użyciu metod Rungego-Kutty. Wyniki porównania wskazały na nieakceptowalne błędy modeli stacjonarnych w sytuacjach zbliżonych do rzeczywistych. Podobne wyniki, uzyskane przez porównanie stacjonarnych przybliżeń bardziej zaawansowanych Markowskich kolejkowych modeli narodzin i śmierci, rozszerzających model Erang-C m.in. o możliwość opuszczenie kolejki bez obsługi lub równoczesne wykonywanie połączeń wychodzących, z wynikami symulacji Monte Carlo, podaje (Deslauriers et al., 2007). W szczególności dotyczy to sytuacji dużej zmienności modelu w czasie. Ingolfsson w (Ingolfsson et al., 2007) dokonał porównania sześciu proponowanych w literaturze metod obliczania zmiennych w czasie modeli call center uzyskując najlepsze rezultaty dla metody uniformizacji dającej te same wyniki co metody Rungego-Kutty przy znacząco niższych nakładach obliczeniowych. W (Ingolfsson et al., 2010) autorzy zaproponowali metodę planowania obsady wykorzystującą uniformizację dla zmiennego w czasie modelu $M_t/M/S_t$, którą porównali z proponowanymi w literaturze metodami wykorzystującymi różne przybliżenia stacjonarne tego samego modelu kolejkowego. Uzyskane wyniki pozwalały uniknąć naruszania ograniczeń jakości serwisu w wyniku błędów przybliżeń stacjonarnych, przy jednoczesnej redukcji kosztów osobowych wygenerowanych grafików obsady o ok. 11%. Niestety nakład obliczeniowy metody pozwalał na jej praktyczne zastosowanie jedynie dla relatywnie małych systemów (poniżej ok. 100 agentów), m.in. ze względu na złożoność obliczeniową wyznaczania funkcji wynikowej (service level) w algorytmie optymalizacji obsady wynoszącą $O(K^2)$ w funkcji wielkości systemu. W artykule (Creemers, Defraeye, and Nieuwenhuyse, 2014) autorzy wykorzystują uniformizację dla obliczania stanów nieustalonych niejednorodnego w czasie Markowskiego modelu czasu ciągłego, w którym odwzorowują inne niż Markowskie dystrybucje czasów obsługi i cierpliwości (w odniesieniu do procesu opuszczania systemu przed obsługa). Możliwe do modelowania wielkości call center są jednak również ograniczone, wymaganym nakładem obliczeniowym, do jedynie niewielkich systemów.

Inne proponowane metody modelowania zmiennych w czasie systemów kolejkowych, to symulacje Monte Carlo, które pozwalają na praktycznie dowolne odwzorowanie cech modelowanego systemu – niestety, nakład obliczeniowy (czas symulacji) potrzebny do uzyskania zadowalającej dokładności jest wielokrotnie większy niż przy metodach wykorzystujących modele CTMC. Przykładami wykorzystania symulacji Monte Carlo do optymalizacji obsady są m.in. (Cezik

and L'Ecuyer, 2008) lub (Feldman et al., 2008). Inną proponowaną metodą jest aproksymacja dyfuzyjna (eng. *fluid/diffusion approximation*), którą wyróżnia niski nakład obliczeniowy, pratycznie niezależny od wielkości systemu (wykorzystywana do planowania obsady call center m.in. w Henken, 2007). Dokładność tej metody rośnie wraz z wielkością systemu, zwłaszcza w przypadku systemów przeciążonych (Whitt, 2006 lub Czachórski et al., 2009). Może być jednak ona niezadowalająca – dla jedynie sporadycznie przeciążonych zmiennych w czasie systemów kolejkowych o wysokim service level – typowych dla większości call center. Wyczerpujące porównanie metod proponowanych do modelowania zmienności w czasie dla systemów obsługi można znaleźć m.in. w (Green, Kolesar, and Whitt, 2007).

Rozdział 2

Metody naukowe stosowane podczas wykonania badań

Rozprawa formułuje problem modelowania rzeczywistych systemów obsługi przy użyciu matematycznej reprezentacji Markowskich modeli kolejkowych jako niejednorodnych w czasie łańcuchów Markowa z czasem ciągłym.

W pracy dokonano analizy stosowanych obecnie algorytmów służących do modelowania takich systemów pod kątem ich wydajności obliczeniowej i możliwości jej poprawy, przy wykorzystaniu cech modelowanych systemów, jak również ścisłego ograniczenia błędu uzyskanego wyniku. Zaproponowane w wyniku tej analizy zmiany algorytmu uniformizacji z wykrywaniem stacjonarności pozwalają na znaczącą poprawę jego efektywności, dla praktycznie używanych modeli rzeczywistych systemów obsługi, oraz ścisłe ograniczenie błędu rozwiązania, co zostało udowodnione w sposób formalny za pomocą matematycznej analizy.

Praca wykorzystuje studia przypadku dla oceny dokładności i efektywności numerycznych metod obliczania stanów nieustalonych ww. modeli. W szczególności, wykorzystano proponowany w literaturze zestaw danych rzeczywistych w celu weryfikacji możliwej do osiągnięcia poprawy wydajności obliczeniowej wynikającej z zaproponowanych zmian.

Rozdział 3

Główny cel rozprawy

Głównym celem pracy jest opracowanie i implementacja rozwiązania pozwalającego na modelowanie obsługi klienta w call center uwzględniającego zmienność modelowanego systemu w czasie, z wydajnością pozwalającą na jego praktyczne zastosowanie do planowania i optymalizacji obsady. W szczególności powinno ono pozwalać na odwzorowanie stanów nieustalonych wynikających ze zmian (natężenie ruchu, czas obsługi, obsada) realistycznego modelu uwzględniającego m.in. ograniczoną pojemność systemu, możliwość jego przeciążenia oraz możliwość rezygnacji klienta z obsługi.

Dodatkowym celem pracy jest weryfikacja osiągniętej poprawy wydajności przy wykorzystaniu danych rzeczywistych - w szczególności w zależności od wielkości i obciążenia modelowanego systemu.

Rozdział 4

Zadania do rozwiązywania

Praca obejmuje następujące zadania główne:

Implementację zaproponowanego w pracy algorytmu w języku C++ umożliwiającą wydajne modelowanie rzeczywistych call center przy użyciu wybranych proponowanych w literaturze Markowskich kolejkowych modeli narodzin i śmierci, uwzględniających stany nieustalone wynikające ze zmienności modelu w czasie.

Przetestowanie działania algorytmu przy wykorzystaniu standardowych modeli Erlang-C i jego rozszerzenia Palm/Erlang-A, jak również proponowanych w literaturze zależnych od stanu modeli pozwalających na lepsze odwzorowanie rozkładów procesu opuszczania kolejki bez obsługi - w szczególności rozszerzenie modeli Erlang-C i Palm/Erlang-A o możliwość natychmiastowego opuszczenia kolejki, po otrzymaniu informacji o konieczności oczekiwania (ang. *balking*).

Przetestowanie teoretycznych założeń algorytmu dotyczących wzrostu wydajności i ograniczenia błędu rozwiązania, na przykładach symulujących skokowe zmiany obsady oraz ciągłe w czasie zmiany intensywności procesu żądań obsługi dla typowych sytuacji obciążenia (przeciążenia) systemu. Przetestowanie wpływu założeń poszczególnych, proponowanych w literaturze modeli na poprawę wydajności w wyniku zaproponowanych zmian algorytmu uniformizacji z wykrywaniem stacjonarności.

Weryfikację możliwego do uzyskania w praktyce polepszenia wydajności obliczania wybranych modeli, przy uwzględnieniu wszystkich proponowanych modyfikacji algorytmu, dla rzeczywistych danych średniej wielkości call center.

Rozdział 5

Wartość teoretyczna

Praca zawiera propozycję modyfikacji szeroko stosowanego algorytmu uniformizacji z wykrywaniem stacjonarności zaproponowanego przez Trivediego i jego współpracowników (Muppala and Trivedi, 1992). Modyfikacja pozwala na uniknięcie błędu przedwczesnego wykrycia stacjonarności analizowanego wcześniej m.in. przez zespół Katoena z RTWE Aachen (Katoen and Zapreev, 2006, Katoen et al., 2011, Zapreev, 2008) i sygnalizowanego również przez inne zespoły zajmujące się wykorzystaniem modeli CTMC (np. Younes et al., 2006). Modyfikacja analizuje numerycznie funkcję konwergencji podporządkowanego łańcucha Markowa czasu dyskretnego w algorytmie uniformizacji, w celu określenia przewidywanego błędu zastąpienia rozwiązania nieustalonego jego stacjonarnym przybliżeniem, wyznaczonym z góry (przy użyciu równań równowagi szczegółowej w przypadku Markowskich kolejkowych modeli narodzin i śmierci). Modyfikacja pozwala równocześnie na wzrost efektywności dla modeli, których rozkład stacjonarny może być obliczony z wydajnością lepszą niż metoda potęgowa (ang. *power method*) – dla liczby iteracji odpowiadającej dolnemu ograniczeniu (ang. *left truncation point*) dystrybuanty dyskretnego rozkładu Poissona, określającego aktywność modelu w badanym okresie czasu. Dodatkowo zaproponowana heurystyka redukcji stanów, wykorzystująca również wyznaczone z góry rozwiązanie stacjonarne, pozwala na zmniejszenie złożoności obliczeniowej w funkcji wielkości systemu, przy jednoczesnym ścisłym ograniczeniu wzrostu błędu rozwiązania do z góry założonej wartości.

Rozdział 6

Wartość praktyczna

Zaproponowane modyfikacje algorytmu uniformizacji z wykrywaniem stacjonarności pozwalają na znaczącą poprawę wydajności obliczeniowej (>100 razy dla średniej wielkości modelu dla danych rzeczywistych), co w połączeniu ze zmniejszeniem złożoności obliczeniowej w funkcji wielkości systemu z $O(K^2)$ do ok. $O(K^{1.3})$ umożliwia użycie go do planowania i optymalizacji obsady dla praktycznie dowolnych, spotykanych w praktyce wielkości call center.

Rozdział 7

Akceptacja wyników przez społeczność naukową

Wyniki pracy były prezentowane na następujących konferencjach:

International Interdysciplinary PhD Workshop - Międzyzdroje, Maj 2015.
Computer Networks CN 2015 - Brunów, Czerwiec 2015.
Computer Networks CN 2016 - Brunów, Czerwiec 2016.

w trakcje recenzji znajduje się obecnie również publikacja planowana do prezentacji na:
International Conference on Methods and Models in Automation and Robotics (MMAR) - Międzyzdroje, Sierpień 2018.

Cząstkowe wyniki pracy zostały opublikowane jako następujące, recenzowane publikacje:

Burak M.:*Inhomogeneous CTMC Model of a Call Center with Balking and Abandonment*, Studia Informatica, vol. 36, nr 2, 2015, s. 23–34,
Burak M.:*Performance analysis of an inbound call center with time varying arrivals*, Zeszyty Naukowe Uniwersytetu Szczecińskiego. Service Management, vol. 15, 2015, s. 5–11,
Burak M.:*Application of an inhomogeneous CTMC model for a telephone call center*, Studia Informatica, vol. 37, nr 2, 2016, s. 15–27,
Burak M.:*Computing discrete Poisson probabilities for uniformization algorithm*, Studia Informatica, vol. 38, nr 1B, 2017, s. 77–88,
Burak M., Korytkowski P.:*Inhomogeneous CTMC Models of Birth-and-Death Systems Solved by Uniformization with Steady-State Detection (in review)*, INFORMS Journal on Computing, 2018,

Rozdział 8

Osiągnięcia zgłaszane w ramach obrony pracy

Praca przedstawia modyfikację szeroko wykorzystywanego wariantu algorytmu uniformizacji z wykrywaniem stacjonarności opublikowanego pierwotnie przez Trivediego i jego współpracowników (Muppala and Trivedi, 1992). Proponowana modyfikacja daje szczególnie korzystne wyniki dla niejednorodnych w czasie łańcuchów Markowa z czasem ciągłym, których stan jest bliski ustalonemu przez znaczącą część analizowanego okresu czasu i posiadają pojedynczy stan ustalony, który może być wyznaczony metodą o wydajności lepszej niż metoda potęgowa, na przykład – dla kolejkowych Markowskich modeli narodzin i śmierci – przy wykorzystaniu równań równowagi szczegółowej. Proponowana modyfikacja znacząco zwiększa jego wydajność, przy jednoczesnym ścisłym ograniczeniu błędu rozwiązania, dzięki uniknięciu problemu przedwczesnego wykrywania stacjonarności obecnego w dotychczas proponowanych implementacjach w.w. algorytmu.

Dodatkowo, zaproponowano modyfikacje zmniejszające liczbę niezbędnych operacji mnożenia wektor macierz reprezentujących zasadniczą część nakładu obliczeniowego algorytmu uniformizacji, jak również nakład obliczeniowy pojedynczej operacji mnożenia wektora i macierzy. W szczególności, wykorzystane zostały specyficzne właściwości Markowskich kolejkowych modeli narodzin i śmierci. Dzięki regularnej strukturze macierzy intensywności przejść takich modeli, możliwe było uniknięcie jej przechowywania w pamięci, co skutkuje 5-6 krotnym przyspieszeniem pojedynczej operacji mnożenia wektora i macierzy.

Dodatkowo, zastosowano metodę redukcji stanów wykorzystującą ponownie metodę obliczania rozkładów stacjonarnych, użytą w modyfikacji algorytmu Mupali i Trivediego, która dzięki równoczesnemu zmniejszeniu nakładu pojedynczej operacji mnożenia wektora i macierzy, jak również niezbędnej ich liczby, zmniejsza również złożoność obliczeniową algorytmu w funkcji wielkości modelowanego systemu z $O(K^2)$ na $O(K^{1.3})$.

Przewidywane korzyści wynikające z zastosowanych modyfikacji algorytmu zostały przetestowane na przykładach numerycznych symulujących typowe sytuacje dla zmiennych w czasie modeli rzeczywistych call center: jednego dla skokowych dyskretnych zmian modelu wynikających ze zmian obsady oraz drugiego symulującego wpływ ciągłych w czasie zmian (o różnej prędkości) intensywności procesu generującego żądania obsługi. Ponadto, całościowy wpływ zastosowanych zmian na efektywność proponowanego rozwiązania został przetestowany przy użyciu danych rzeczywistego call center amerykańskiego banku średniej wielkości.

W porównaniu z dotychczas prezentowanymi w literaturze metodami wykorzystującymi niejednorodne modele CTMC, które były ograniczone jedynie do

modelu M/M/S, proponowana metoda umożliwia łatwą implementację dowolnych Markowskich kolejkowych modeli narodzin i śmierci, w tym również modeli zależnych od stanu, pozwalających na odwzorowanie innych niż Markowskie rozkładów czasu obsługi lub czasu cierpliwości procesu opuszczania systemu bez obsługi (ang. *abandonment*).

Rozdział 9

Skrótowa prezentacja struktury i układu pracy

Pierwszy rozdział pracy zawiera wprowadzenie do tematyki. Omawiane są: czym jest call center, jak można opisać podstawowy proces operacyjny call center obsługującego ruch przychodzący, w szczególności przy wykorzystaniu modeli kolejkowych. Opisywane jest również znaczenie modeli dla zarządzania operacyjnego, w szczególności możliwość odwzorowania ilościowych parametrów oceny jakości obsługi (ang. *service level*) będących podstawowym miernikiem efektywności pracy organizacji, używanym m.in. do weryfikacji zapisów service level agreement, umów o świadczeniu usług. Pokazane są przykłady typowych takich parametrów oceny (ang. *key performance indicators - KPI*) jak również sposoby ich obliczania przy wykorzystaniu mierzalnych parametrów (ang. *performance measures*) rzeczywistych call center, bądź też wyników modelowania. Następnie przedstawiony jest problem planowania obsady, który, ze względu na znaczenie kosztów osobowych, jest kluczowy dla efektywności organizacji.

Rozdział drugi omawia zastosowanie kolejkowych modeli call center. Na wstępie przedstawione jest krótkie wprowadzenie do reprezentacji Markowskich modeli kolejkowych jako łańcuchy Markowa z czasem dyskretnym (DTMCs) i ciągłym (CTMCs), w tym również problem istnienia i osiągania stanu ustalonego dla łańcuchów Markowa z czasem dyskretnym wykorzystywany w proponowanej w pracy modyfikacji algorytmu uniformizacji. Następnie przedstawiana jest klasa Markowskich modeli narodzin i śmierci na przykładzie standardowych modeli Erlang-B i Erlang-C. W dalszej części analizowana jest problematyka adekwatności odwzorowania rzeczywistych call center za pomocą modeli proponowanych w literaturze. Jako punkt wyjścia wybrano powszechnie stosowany w praktyce stacjonarny model $M/M/S$ (Erlang-C). Analizowana jest jego, wynikająca z przyjętych założeń, dokładność, jak również proponowane w literaturze modyfikacje, uwzględniające takie cechy systemów rzeczywistych jak m.in. ich niejednorodność w czasie, konieczność odwzorowania abandonment, jak również innych niż Markowskie rozkładów procesu obsługi. Dla wybranych modeli, wykorzystywanych w dalszej części pracy jako przykładowe do analizy wydajności obliczeniowej proponowanych algorytmów, podane są również sposoby wyznaczania typowych wykorzystywanych w praktyce performance measures. Dotyczy to, w szczególności, uwzględniającego abandonment modelu $M/M/S/K+M$ (Palm/Erlang-A) jak również przykładów $M(n)/M/S/K$ oraz $M(n)/M/S/K+M$ modeli zależnych od stanu. Na zakończenie porównywane są metody optymalizacji obsady z wykorzystaniem modeli kolejkowych, wykorzystujące zarówno ich przybliżenia stacjonarne, jak i metody uwzględniające stany nieustalone modelu wynikające z jego zmienności w czasie.

Rozdział trzeci jest poświęcony modyfikacjom algorytmu uniformizacji z wykrywaniem stacjonarności, zastosowanym dla niejednorodnych w czasie łańcuchów Markowa czasu ciągłego, które są głównym osiągnięciem pracy. Punktem wyjścia jest oryginalny algorytm rozwiązujący jednorodne CTMCs, zaproponowany przez Jensena w (Jensen, 1953) i jego rozszerzenie o wykrywanie stacjonarności podporządkowanego DTMC zaproponowane w (Muppala and Trivedi, 1992), dla którego analizowane są jego złożoność obliczeniową i ograniczenie błędu rozwiązania. Następnie, bazując na właściwościach zbieżności ww. podporządkowanego DTMC do stanu ustalonego, wyprowadzona zostaje propozycja modyfikacji, poprawiającą efektywność algorytmu, przy jednoczesnym uniknięciu tzw. przedwczesnego wykrycia stacjonarności, wraz z analizą otrzymanego ścisłego ograniczenia błędu rozwiązania. Następnie przedstawione jest jego zastosowanie dla niejednorodnych CTMCs, wraz z analizą błędu rozwiązania. Finalnie, proponowane są dodatkowe usprawnienia algorytmu, bazujące na szczególnych właściwościach CTMCs, reprezentujących Markowskie kolejkowe modele narodzin i śmierci – w szczególności metoda redukcji stanów.

W rozdziale czwartym dysertacji prezentowane są wyniki przeprowadzonych dwóch, modelujących reprezentatywne dla rzeczywistych call center sytuacje, eksperymentów numerycznych: pokazujących wpływ dyskretnych (skokowych) zmian obsady, oraz zmieniającej się w sposób ciągły intensywności procesu przychodzących żądań obsługi. Na przykładzie o stałej intensywności strumienia zgłoszeń pokazany został wpływ rodzaju zastosowanego modelu na szybkość zbieżności do stanu ustalonego, po skokowej zmianie liczby serwerów, i wynikające z tego oszczędności czasu obliczeń (liczby niezbędnych operacji mnożenia wektor macierz) dzięki możliwości aproksymacji stanu nieustalonego rozkładem stacjonarnym, dla założonego z góry błędu takiego przybliżenia. Dodatkowo, zweryfikowany został błąd rozwiązania, poprzez porównanie z wynikami obliczeń bez wykorzystania wykrywania stacjonarności. Przykład o stałej liczbie serwerów został wykorzystany dodatkowo do pokazania wpływu zmienności intensywności strumienia zgłoszeń, średniego czasu obsługi (poprzez porównanie okresów nieprzeciążonych z okresami przeciążonymi, w których rolę procesu "obsługi" przejmuje proces abandonment) oraz wielkości systemu, na możliwą do uzyskania redukcję nakładu obliczeniowego, dzięki wykorzystaniu proponowanych modyfikacji algorytmu uniformizacji. Dodatkowo zbadany został wpływ zaproponowanej heurystyki redukcji stanów na złożoność obliczeniową w funkcji wielkości systemu.

Rozdział piąty wykorzystuje dane call center średniej wielkości amerykańskiego banku w celu weryfikacji możliwej do uzyskania redukcji nakładu obliczeniowego w sytuacjach rzeczywistych. W szczególności użyty jest zaproponowany w aktualnej literaturze zbiór danych jednego miesiąca, w celu skonstruowania przykładowej prognozy ruchu i weryfikacji wyników modelu dla przykładowej obsady – analogicznie do procedury wykorzystywanej w metodzie optymalizacji obsady zaproponowanej przez Ingolfsson et al., 2010.

Rozdział szósty zawiera podsumowanie oraz proponowane kierunki dalszych badań.

Część pracy, w szczególności w rozdziałach 3, 4 i 5, wykorzystuje własne publikacje autora: Burak, 2014, Burak, 2015b, Burak, 2015a, Burak, 2016, Burak and Korytkowski, 2018, zgodnie z zasadami udostępniania publikowanych treści ich wydawców.

Rozdział 10

Zawartość pracy

W tym rozdziale omówione są proponowane modyfikacje algorytmu uniformizacji dla modelowania cal center, zaproponowane w rozdziale trzecim dysertacji, które są głównym osiągnięciem pracy. W pierwszej kolejności zostanie omówiony oryginalny algorytm uniformizacji dla jednorodnych łańcuchów Markowa z czasem ciągłym, jak również jego modyfikacja wykorzystująca wykrywanie stacjonarności, ze szczególnym uwzględnieniem złożoności obliczeniowej oraz kontroli błędów. Następnie przedstawione zostają propozycje modyfikacji ww. algorytmu wykorzystujące właściwości funkcji zbieżności analizowanego modelu, jak również ich zastosowanie dla niejednorodnych CTMCs. Dodatkowo prezentowana jest heurystyka redukcji stanów, dla kolejkowych Markowskich modeli narodzin i śmierci, obniżającą złożoność obliczeniową algorytmu w funkcji wielkości analizowanego systemu. Następnie opisane są eksperymenty numeryczne przeprowadzone w celu weryfikacji przewidywanych rezultatów naszych propozycji.

10.1 Metoda uniformizacji dla jednorodnych CTMCs

Algorytm uniformizacji nazywany również randomizacją lub algorytmem Jensena oferuje stabilną obliczeniowo i wydajną metodę wyznaczania stanów nieustalonych jednorodnych CTMCs. Jego zaletą jest możliwość ograniczenia błędu rozwiązania do z góry założonej wartości jak również wydajność wyraźnie przewyższająca inne metody numerycznego rozwiązywania układów równań różniczkowych (np. metody Rungego-Kutty), co zostało pokazane dla różnych analizowanych modeli m.in. w (Grassmann, 1978), (Reibman and Trivedi, 1988), (Arns, Buchholz, and Panchenko, 2010) lub dla modelu call center w (Ingolfsson et al., 2007). Jest on również standardowym algorytmem używanym przez pakiety do analizy modeli opisanych przez CTMCs (m.in. PRISM – Kwiatkowska, Norman, and Parker, 2017).

Jednorodny CTMC może być opisany jednoznacznie za pomocą *generatora infinitezymalnego* (także: *macierzy tranzycji*, *macierzy intensywności przejść*) Q zawierającego intensywności przejść q_{ij} z dowolnego stanu i do stanu j jak również początkowego wektora prawdopodobieństw stanów $\pi(0)$. Prawdopodobieństwa jego zależnych od czasu (nieustalonych) stanów $\pi(t) = [\pi_0(t), \ldots, \pi_K(t)]$, gdzie $\pi_k(t)$ oznacza prawdopodobieństwo systemu znajdującego się w stanie k,

mogą być opisane przez następujący system równań różniczkowych:

$$\pi'(t) = \pi(t)Q \tag{10.1}$$

Jego rozwiązanie za pomocą uniformizacji wymaga w pierwszej kolejności utworzenia macierzy

$$P = I + \frac{Q}{\alpha} \tag{10.2}$$

która dla *współczynnika uniformizacji* $\alpha \geq max_i(|q_{i,i}|)$ jest macierzą stochastyczną. Dodatkowo, przyjmując

$$\beta(\alpha t, i) = e^{-\alpha t}\frac{(\alpha t)^i}{i!} \tag{10.3}$$

jako prawdopodobieństwo wygenerowania dokładnie i zdarzeń w przedziale czasu $[0, t)$ przez dyskretny proces Poissona o intensywności α, otrzymuje się

$$\pi(t) = \pi(0)\sum_{i=0}^{\infty}\beta(\alpha t, i)(P)^i \tag{10.4}$$

Aby uniknąć pojawiania się nowych niezerowych elementów (ang. *fill-in*) w rzadkiej macierzy P na skutek jej potęgowania, można obliczyć alternatywnie sumę:

$$\pi(t) = \sum_{i=0}^{\infty}\beta(\alpha t, i)(\pi(0)P^i) = \sum_{i=0}^{\infty}\beta(\alpha t, i)\nu(i) \tag{10.5}$$

gdzie $\nu(i)$ odpowiada wektorowi prawdopodobieństw stanów łańcucha Markowa czasu dyskretnego (DTMC) opisanego macierzą prawdopodobieństw przejść P po i ewolucjach (krokach czasu dyskretnego), który może zostać wyznaczony iteracyjnie:

$$\nu(0) = \pi(0), \nu(i) = \nu(i-1)P \tag{10.6}$$

bez konieczności przechowywania w pamięci potęg macierzy P nie będących macierzami rzadkimi.

Równanie 10.5 może być interpretowane w sposób probabilistyczny, jako (dyskretny) proces Markowa z czasem dyskretnym (DTMC) napędzany procesem Poissona generującym zdarzenia z intensywnością α. Ponieważ prawdopodobieństwo zajścia dokładnie i przejść DTMC w przedziale czasu $[0, t)$ jest równe $\beta(\alpha t, i)$ – to, zgodnie z twierdzeniem o prawdopodobieństwie całkowitym, prawdopodobieństwa stanów (CTMC) $\pi(t)$, dla czasu t, będą równe sumie wektorów prawdopodobieństw DTMC odpowiadających wszystkim możliwym liczbom ewolucji (kroków) w $[0, t)$, ważonej prawdopodobieństwem ich wystąpienia, wynikającym z (dyskretnego) rozkładu Poissona.

Aby obliczenie $\pi(t)$ było możliwe w skończonym czasie, oczywista jest konieczność ograniczenia nieskończonego dodawania w 10.5 po k ewolucjach DTMC. Otrzymana w wyniku tego wartość

$$\pi(t) = \sum_{i=0}^{k}\beta(\alpha t, i)\nu(i) \tag{10.7}$$

różni się od dokładnej wartości $\pi(t)_\infty$ – uzyskanej w wyniku nieskończonego wykonywania dodawania – nie więcej niż wartość ogona dystrybuanty rozkładu Poissona:

$$\|\pi(t)_\infty - \pi(t)\| \leq 1 - \sum_{i=0}^{k} e^{-\alpha t} \frac{(\alpha t)^i}{i!} \leq \epsilon \tag{10.8}$$

Analogicznie do propozycji zawartej np. w artykule (Reibman and Trivedi, 1988)) będziemy określać k jako ograniczenie górne (ang. *right truncation point*), którego wartość może być wyznaczona z góry dla zadanego ograniczenia błędu rozwiązania (ang. *error bound*) ϵ.

Wraz z wzrastającym αt będzie również malało prawdopodobieństwo wystąpienia małej liczby zdarzeń w $[0, t)$. Pozwala to na dalsze ograniczenie sumowania tylko dla wektorów prawdopodobieństw stanów DTMC odpowiadającym liczbie ewolucji (zdarzeń generowanych przez proces Poissona), których prawdopodobieństwo wystąpienia jest istotne – tzn. rozpoczęcia go od ewolucji l, nazywanej ograniczeniem dolnym (ang. *left truncation point*). Równanie 10.7 redukuje się konsekwentnie do:

$$\pi(t) = \sum_{i=l}^{k} \beta(\alpha t, i)\nu(i) \tag{10.9}$$

gdzie wartości l i k są określone (np. zgodnie z propozycją w Reibman and Trivedi, 1988) jako:

$$\sum_{i=0}^{l-1} e^{-\alpha t} \frac{(\alpha t)^i}{i!} \leq \frac{\epsilon}{2},\ 1 - \sum_{i=0}^{k} e^{-\alpha t} \frac{(\alpha t)^i}{i!} \leq \frac{\epsilon}{2} \tag{10.10}$$

Wartości ograniczeń l i k jak również niezbędne wartości prawdopodobieństw wystąpienia $l \leq n \leq k$ zdarzeń wygenerowanych przez proces Poissona, mogą być wyznaczone z góry np. przy użyciu algorytmu zaproponowanego w Burak, 2017 – będącego rozwinięciem popularnego algorytmu Foxa-Glynna (Fox and Glynn, 1988).

10.1.1 Złożoność obliczeniowa

Podstawowym czynnikiem determinującym złożoność obliczeniową algorytmu uniformizacji jest konieczność obliczenia k ewolucji podporządkowanego DTMC tzn. dokonania k operacji mnożenia wektora $\nu(i)$ $(0 \leq i \leq k)$ i macierzy P o rozmiarze $K + 1$ gdzie K oznacz wielkość systemu. W przypadku Markowskich kolejkowych modeli narodzin i śmierci macierz P jest trójdiagonalna, w związku z tym zarówno złożoność obliczeniowa mnożenia wektor macierz (MVM) jak i zapotrzebowanie na pamięć zależą liniowo od wielkości systemu. Wraz ze wzrostem αt rozkład Poissona, będzie, zgodnie z centralnym twierdzeniem granicznym, zbieżny do rozkładu normalnego. W związku z tym można założyć, że dolne i górne ograniczenia l i k będą asymptotycznie symetryczne do wartości oczekiwanej. Liczba ewolucji będzie w związku z tym wynosiła $\frac{l+k}{2}$ (zależne liniowo od αt) powiększona o dodatkowe $\frac{k-l}{2}$ ewolucji (MVM) koniecznych do uzyskania zadanej dokładności, których liczba jest zależna od pierwiastka wartości oczekiwanej (tzn. $O\sqrt{\alpha t}$) i wymaganej dokładności ϵ. W szczególności koszt uzyskania wyższej dokładności rozwiązania maleje proporcjonalnie z αt (wysoka

aktywność modelu lub długie przedziały czasu). Dla modeli, w których liczba serwerów jest proporcjonalna do wielkości systemu oraz takich, które modelują proces abandonment o wykładniczym rozkładzie cierpliwości (np. Palm/Erlang-A i jego rozszerzenia), współczynnik uniformizacji α będzie rósł liniowo wraz ze wzrostem K – w związku z tym całkowita złożoność obliczeniowa będzie $O(K^2t)$. Wyższe wartości l powodują zmniejszenie liczby wektorów $\nu(i)$ $(l \leq i \leq k)$ biorących udział w ważonej (prawdopodobieństwami wynikającymi z dyskretnego rozkładu Poissona) sumie – oszczędność z tego wynikająca jest jednak nieznaczna za wyjątkiem sytuacji, kiedy możliwe byłoby wyznaczenie $\nu(l)$ w sposób bardziej wydajny niż za pomocą metody potęgowej.

10.1.2 Uniformizacja z wykrywaniem stacjonarności

Równanie 10.6 jest równoważne metodzie potęgowej wyznaczania ewolucji DTMC. Jeżeli DTMC opisany macierzą P jest nieprzywiedlny i nieokresowy, to będzie posiadał jedno rozwiązanie ustalone (stacjonarne), niezależne od początkowego wektora prawdopodobieństw stanów $\nu(0) = \pi(0)$, do którego 10.6 będzie zbieżne. W przypadku gdyby DTMC opisany przez P osiągnął stan stacjonarny przed dolnym ograniczeniem l dalsze operacje mnożenia wektor macierz nie byłyby potrzebne. Wykorzystanie tego faktu w algorytmie uniformizacji zostało zaproponowane pierwotnie przez Muppalę i Trivediago w (Muppala and Trivedi, 1992). Proponowana przez nich modyfikacja zakłada, że DTMC opisany przez P ma jedno ustalone rozwiązanie $\nu(\infty)$ i że po S ewolucjach 10.6 $\|\nu(\infty) - \nu(S)\|_v < \delta(S)$, gdzie $\|.\|_v$ jest dowolnie wybraną normą. Równanie 10.9 przybiera wtedy formę:

$$\hat{\pi}(t) = \begin{cases} \nu(S) & \text{jeżeli } S \leq l, \\ \sum_{i=l}^{S} \nu(i) e^{-\alpha t} \frac{(\alpha t)^i}{i!} + \nu(S)(1 - \sum_{i=0}^{S} e^{-\alpha t} \frac{(\alpha t)^i}{i!}) & \text{jeżeli } l < S \leq k, \\ \text{tak jak } \pi(t) \text{ w (10.9)} & \text{jeżeli } S > k \end{cases} \tag{10.11}$$

gdzie $\hat{\pi}(t)$ użyte zamiast of $\pi(t)$ oznacza aproksymację nieustalonego stanu przez wykryty stacjonarny wektor prawdopodobieństw stanu $\nu(S)$.

W celu zapewnienia tego samego ograniczenia błędu rozwiązania ϵ w przypadkach kiedy stacjonarność została wykryta, w (Malhotra, Muppala, and Trivedi, 1994) zostało zaproponowane zachowanie następującej nierówności:

$$\|\pi(t) - \hat{\pi}(t)\| < \frac{\epsilon}{2} + 2\delta(S) \tag{10.12}$$

Przedstawiona przez nich analiza błędu zakłada jednak, że błąd uzyskanego (wykrytego) stanu stacjonarnego w odniesieniu do dokładnej wartości stanu ustalonego jest znany (mniejszy od założonego ograniczenia błędu) – dyskusję ich propozycji można znaleźć m.in. w (Younes et al., 2006).

10.2 Zmodyfikowany algorytm uniformizacji z wykrywaniem stacjonarności

Jeżeli zbieżność metody potęgowej jest zapewniona, to za (Stewart, 2009), liczba jej iteracji k niezbędna do uzyskania dokładności rozwiązania lepszej niż ξ może

być wyznaczona w przybliżeniu z zależności:

$$\rho^k \approx \xi, \text{ tzn., } k \approx \frac{log\xi}{log\rho} \tag{10.13}$$

gdzie ρ jest rzędem wielkości wartości własnej λ_2 macierzy P

$$1 = \|\lambda_1\| > \|\lambda_2\| \geq \|\lambda_3\| ... \geq \|\lambda_N\| \tag{10.14}$$

Złożoność obliczeniowa algorytmu uniformizacji w przypadku wykorzystania wykrycia stacjonarności wynosi konsekwentnie $O(\eta\ log\xi/log|\lambda_2|)$. Jak wskazano w Katoen and Zapreev, 2006 powyższa zależność zależy od zbieżności postępującej geometrycznie co wymaga użycia normy maksimum (ang. *total variation norm*) l^∞, zdefiniowanej jako $\|\nu\|_\infty = max_i|\nu_i|$.

Ponieważ wartość własna macierzy nie jest na ogół znana (łatwa do obliczenia z góry) zwyczajowo wykrycie zbieżności następuje poprzez wyznaczenie normy różnicy wektorów kolejnych iteracji:

$$\|\nu_i(k) - \nu_i(k-m)\| < \xi \tag{10.15}$$

Na przykład Stewart, 2009 proponuje testowanie różnicy wyników co m iteracji, gdzie m powinno zależeć od szybkości zbieżności:

$$max_i \left(\frac{|\nu_i(k) - \nu_i(k-m)|}{|\nu_i(k)|} \right) < \xi \tag{10.16}$$

i sugeruje zastosowanie dodatkowej "baterii testów" w celu zapewnienia, że przybliżenie $\nu(S)$ będzie wystarczająco dokładne.

Zasadniczym problemem tej propozycji jest nakład potrzebny do zapewnienia, że $\nu(S)$ jest ustalony, czyli że błąd $\delta(S)$ może być traktowany jako pomijalny. Dodatkowo może nastąpić tzw. przedwczesne wykrycie stacjonarności prowadzące do znaczących błędów takiej aproksymacji omawiane na przykład w (Katoen and Zapreev, 2006) i (Younes et al., 2006). Według (Bolch et al., 2006) nie jest również możliwe określenie z góry punktu, w którym $\nu(i)$ będzie bliski stanowi ustalonemu z góry założonym marginesem błędu.

W pewnych przypadkach, możliwe jest jednak wykorzystanie faktu, że istnieje łatwy sposób wyznaczenia z góry dokładnej wartości rozkładu stacjonarnego. W szczególności w przypadku Markowskich kolejkowych modeli narodzin i śmierci możemy użyć w tym celu równań równowagi szczegółowej (np. w sposób opisany w Stewart, 2009 lub Ingolfsson and Tang, 2012). W związku z tym – zamiast obliczać kolejne iteracje DTMC w (10.6) do momentu aż będzie można w uzasadniony sposób przyjąć, że $\nu(S)$ jest ustalony (np. wykonując proponowane w literaturze testy na zbieżność) –można, jak zaproponowano pierwotnie w Burak, 2014, użyć dokładnej, wyznaczonej z góry wartości $\nu(\infty)$ (zamiast $\nu(S)$, proponowanego w Muppala and Trivedi, 1992) jako przybliżenie $\hat{\pi}(t)$ rzeczywistego $\pi(t)$, w sytuacji, kiedy błąd takiego przybliżenia, który jest obecnie dokładnie znany, jest dla pewnej iteracji s w (10.6) (gdzie $s < l$ dla l przy ϵ_Δ jak w (10.9)) mniejszy od założonego ograniczenia błędu rozwiązania:

$$\frac{\|\nu(s) - \nu(\infty)\|_\infty}{\|\nu(\infty)\|_\infty} < \delta \tag{10.17}$$

Pozwala to na pominięcie kolejnych iteracji DTMC $\nu(i), i > s$ w celu obliczenia $\pi(t)$ jak w (10.5), dzięki użyciu $\nu(\infty)$ jako przybliżenia $\hat{\pi}(t)$ dla rozwiązania $\pi(t)$:

$$\hat{\pi}(t) = \begin{cases} \nu(\infty) & \text{jeżeli } s \leq l, \\ \sum_{i=l}^{s} \nu(i) e^{-\alpha t} \frac{(\alpha t)^i}{i!} + \nu(\infty)(1 - \sum_{i=0}^{s} e^{-\alpha t} \frac{(\alpha t)^i}{i!}) & \text{jeżeli } l < s \leq k, \\ \text{tak jak } \pi(t) \text{ w (10.9)} & \text{jeżeli } s > k \end{cases} \tag{10.18}$$

Różnica pomiędzy tak uzyskanym przybliżeniem a dokładną wartością $\pi(t)_\infty$ otrzymaną przy założeniu braku ograniczeń sumowania (tzn. dla $l = 0$ i $k = \infty$) jest równa:

$$\begin{aligned} \pi_\infty(t) - \hat{\pi}(t) &= \sum_{i=0}^{s} (\nu(i) - \nu(\infty)) e^{-\alpha t} \frac{(\alpha t)^i}{i!} \\ &+ \sum_{i=s+1}^{\infty} (\nu(i) - \nu(\infty)) e^{-\alpha t} \frac{(\alpha t)^i}{i!} \end{aligned} \tag{10.19}$$

która, w związku z tym że $\nu(i)$ i $\nu(\infty)$ są wektorami stochastycznymi, a co za tym idzie $\forall : j \ |\nu_j(i) - \nu_j(\infty)| < 1$ oraz dla $s < l$, norma pierwszej sumy jest ściśle ograniczona przez:

$$\epsilon_s = Q_\lambda(s) = \sum_{i=0}^{s} e^{-\alpha t} \frac{(\alpha t)^i}{i!} \quad , \epsilon_s < \frac{\epsilon_\Delta}{2} \tag{10.20}$$

Dodatkowo, z pierwotnego założenia wynika że:

$$\forall i > s : \|\nu(i) - \nu(\infty)\|_\infty < \delta \|\nu(\infty)\|_\infty, \text{ i } \sum_{i=s+1}^{\infty} e^{-\alpha t} \frac{(\alpha t)^i}{i!} < 1$$

Wynikający błąd aproksymacji $\pi(t)$ przez $\nu(\infty)$ jest, konsekwentnie, ograniczony do:

$$\|\pi(t) - \hat{\pi}(t)\|_\infty < \epsilon_s + \delta \|\nu(\infty)\|_\infty \quad , \epsilon_s < \frac{\epsilon_\Delta}{2} \tag{10.21}$$

Jak wskazano w 10.13 szybkość zbieżności $\nu(i)$ do $\nu(\infty)$, zdefiniowana jako zmiana logarytmu błędu (normy różnicy obu wektorów), będzie dążyła asymptotycznie do stałej wartości zależnej jedynie od wartości własnej λ_2 macierzy P. Niestety wartość ta nie jest na ogół znana, podobnie jak dokładna funkcja szybkości zbieżności w zależności od liczby iteracji $cr(i)$ – co pozwoliłoby ewentualnie na określenie z góry błędu rozwiązania dla dowolnej iteracji $l_S > i$.

Niemniej jednak można, dla dowolnej iteracji i, obliczyć logarytm aktualnego błędu rozwiązania $\frac{\|\nu(i) - \nu(\infty)\|_\infty}{\|\nu(\infty)\|_\infty}$, posługując się wyznaczoną z góry wartością rozkładu ustalonego. Dodatkowo, posługując się jego wcześniejszymi wartościami, można wyznaczyć aktualną wartość funkcji szybkości zbieżności oraz jej

pochodnych. W związku z tym, korzystając z faktu, że szybkość zbieżności dąży asymptotycznie do pewnej stałej wartości (jak w (10.13)) można obliczyć z wyprzedzeniem błąd $\frac{\|\nu(l_S)-\nu(\infty)\|_\infty}{\|\nu(\infty)\|_\infty}$ używając rozwinięcia Taylora:

$$\log(\delta_{l_S}) = log\left(\frac{\|\nu(l_S)-\nu(\infty)\|_\infty}{\|\nu(\infty)\|_\infty}\right) = log\left(\frac{\|\nu(i)-\nu(\infty)\|_\infty}{\|\nu(\infty)\|_\infty}\right) + (l_S - i)cr(i) + \frac{(l_S-i)^2}{2!}cr'(i) \quad (10.22)$$

z rosnącą asymptotycznie dokładnością, jako że $lim_{i\to\infty}(cr'(i)) \to 0$.

Zakładając następnie pewne l_S mniejsze niż dolne ograniczenie l (zdefiniowane jak dla ϵ_Δ w (10.9)), takie że ϵ_{l_S} będzie równe dystrybuancie dyskretnego rozkładu Poissona jak w (10.20) i pomijalnie małe (np. $\epsilon_{l_S} = \epsilon_\Delta \times 10^{-2}$) to – pomimo że dla pewnego $i < l_S$ relatywny błąd $\nu(i)$ jest wyższy niż ograniczenie δ (jak w 10.17) – można o ile $\delta_{l_S} < \delta$ wyznaczony zgodnie z (10.22) zaprzestać dalszego iterowania podporządkowanego DTMC i wykorzystać $\nu(\infty)$ jako $\hat{\pi}(t)$

$$\hat{\pi}(t) = \begin{cases} \nu(\infty) & \text{jeżeli } \delta_{l_S} \leq \delta, \\ \sum_{i=l}^{k} \nu(i) e^{-\alpha t}\frac{(\alpha t)^i}{i!} & \text{if } \delta_{l_S} > \delta \end{cases} \quad (10.23)$$

Błąd uzyskanego rozwiązania będzie ściśle ograniczony zgodnie z (10.21) dla ϵ_{l_S} zastępującego wtedy ϵ_s i δ_{l_S} zastępującego δ do:

$$\|\pi(t) - \hat{\pi}(t)\|_\infty < \epsilon_{l_S} + \delta_{l_S}\|\nu(\infty)\|_\infty \ \ , \epsilon_{l_S} < \frac{\epsilon_\Delta}{2} \quad (10.24)$$

10.2.1 Zastosowanie zmodyfikowanego algorytmu uniformizacji z wykrywaniem stacjonarności dla niejednorodnych w czasie CTMCs

Jedną z najważniejszych właściwości rzeczywistych systemów call center, które należy uwzględnić przy ich modelowaniu jest ich zmienność w czasie. Wynika ona ze zmiennego w funkcji czasu natężenia procesu żądań obsługi co powoduje konieczność zapewnienia odpowiedniej (zmiennej w czasie) liczby serwerów w celu zapewnienia wystarczającej jakości obsługi. Powoduje to niejednorodność (w czasie) łańcucha Markowa czasu ciągłego opisującego wybrany model kolejkowy. Pomimo, że algorytm uniformizacji zakłada jednorodność badanego modelu w analizowanym okresie czasu, możliwe jest również zastosowanie go dla systemów zmiennych w czasie, o ile zmiany w Q zachodzą w sposób dyskretny, zastępując niejednorodny w czasie model szeregiem jednorodnych w poszczególnych odcinkach czasu – jak to zostało zaproponowane m.in. w (Gross and Miller, 1984), (Rindos et al., 1995) or (Arns, Buchholz, and Panchenko, 2010). Należy w tym celu podzielić analizowany odcinek czasu $[0, T)$ na mniejsze odcinki (kroki) o długości Δ. Następnie, dla każdego z takich kroków, zostaje wyznaczony (nieustalony) rozkład prawdopodobieństw na jego koniec wykorzystywany następnie jako początkowy rozkład prawdopodobieństw stanu kolejnego kroku.

Zmiany w Q w funkcji czasu mogą występować w przypadku modelowania call center w sposób dyskretny na skutek zmieniającej się obsady lub w sposób ciągły dla procesu żądań obsługi. W przypadku wykorzystania modelu do optymalizacji obsady, zmiany liczby agentów następują jedynie na początku okresów planowania (o długości na ogół 15 lub 30 minut) a dane prognoz przychodzącego ruchu są zazwyczaj agregowane w nieco krótszych odcinkach czasu (5 do 15 minut). W związku z tym można założyć, podobnie jak m.in. Ingolfsson et al., 2010 dyskretną naturę zmian modelu w czasie.

W przypadku, gdyby dostępne były prognozy ruchu o charakterze ciągłym, możliwe jest wykorzystanie metody zaproponowanej w (Arns, Buchholz, and Panchenko, 2010) umożliwiającej taką dyskretyzacje zmiennego w czasie natężenia ruchu przychodzącego, w której punkty podziału odcinka czasu na kroki są generowane tak, aby błąd powodowany przez taka aproksymację nie przekraczał z góry ustalonej wartości.

Jedną z najważniejszych zalet algorytmu uniformizacji jest możliwość ograniczenia z góry błędu rozwiązania. Dodatkowo, można pokazać (Van Moorsel and Sanders, 1997), że błąd analizy podzielonego na jednorodne odcinki przedziału czasu jest zawsze mniejszy niż suma błędów takich następujących po sobie kroków rozwiązania.

Można sformułować to w sposób następujący: dla odcinka czasu o długości T ze znanym początkowym rozkładem prawdopodobieństw stanów $\pi(0)$, można założyć, że dla każdego $\pi(\tau)$, $\tau = (0, T]$, wartości prawdopodobieństw stanów powinny być wyznaczone z błędem mniejszym niż ε_T. Dalej zakłada się, że błąd po obliczeniu dowolnego $\pi(t), t < T$ wynosi $\varepsilon_t < \varepsilon_T$. Wynika z tego, że:

$$\varepsilon_t + \sum_i \epsilon_{\Delta_i} \leq \varepsilon_T, \ \sum_i \Delta_i = T - t \tag{10.25}$$

W związku z tym górna granica błędu po kroku o długości Δ zaczynającym się od t wyznaczanym z ograniczeniem błędu ϵ_Δ będzie wynosiła:

$$\varepsilon_{t+\Delta} = \varepsilon_t + \epsilon_\Delta \tag{10.26}$$

zgodnie z propozycją w (Arns, Buchholz, and Panchenko, 2010), jeżeli $\varepsilon_R = \varepsilon_T - \varepsilon_t$ jest pozostającym ograniczeniem błędu, to wtedy w kroku o długości $\Delta \leq (T - t)$ z początkowym rozkładem $\pi(t)$ jego ograniczenie błędu powinno wynosić:

$$\epsilon_\Delta \leq \varepsilon_R \frac{\Delta}{T - t} \tag{10.27}$$

aby nie przekroczyć w żadnym momencie analizowanego odcinka czasu T globalnego ograniczenia błędu ε_T. Propozycja ta zakłada więc podział limitu błędu proporcjonalnie do długości kroków. Pomimo, że wydaje się to intuicyjnie właściwe, warto rozważyć, w nawiązaniu do wspomnianej wcześniej złożoności obliczeniowej potrzebnej do uzyskania dodatkowej dokładności wynoszącej asymptotycznie $O(\sqrt{\alpha t})$, większe wartości ograniczenia błędu dla kroków z mniejszym αt (tzn. krótsze odcinki czasu lub mniejsza aktywność modelu). Alternatywnie, można, ograniczając błąd poszczególnych kroków (zwłaszcza w przypadku modeli o dużej wielkości), zwiększyć pozostający do "wykorzystania" globalny limit błędu jako dopuszczalny błąd aproksymacji rozkładem stacjonarnym δ (jak w (10.17)). Jest to możliwe, ponieważ w wypadku w którym DTMC opisany przez

P jest nieprzywiedlny i nieokresowy, to posiada on jedynie jeden stan ustalony niezależny od wartości początkowej $\pi(0)$. W związku z tym błąd aproksymacji rozkładem stacjonarnym dla dowolnego kroku $[t, t+\Delta)$ jest absolutny i niezależny od błędu początkowego (poprzednich kroków):

$$\varepsilon_{t+\Delta} = \epsilon_{l_S} + \delta_{l_S} \|\nu(\infty)\|_\infty \tag{10.28}$$

dla ϵ_{l_S} jak ϵ_s w (10.20) i δ_{l_S} jak w (10.22) i mniejszy niż użyte w zmodyfikowanym algorytmie uniformizacji z wykrywaniem stacjonarności ograniczenie błędu δ aproksymacji rozkładem ustalonym.

Pozwala to na użycie ograniczenia δ zależnego od pozostającej do "wykorzystania" wartości ograniczenia błędu całego rozwiązania a nie od ograniczenia błędu pojedynczego kroku ϵ_Δ jak to zostało zaproponowane w oryginalnym algorytmie w Malhotra, Muppala, and Trivedi, 1994. Zakładając, że model został obliczony do czasu $m, 0 \leq m < T$, to – aby zapewnić $\varepsilon_t < \varepsilon_T$ dla każdego $\pi(t)$, $t = (m, T]$ – powinno być spełnione:

$$\sum_m^T \epsilon_\Delta \leq \varepsilon_T - \varepsilon_m \text{ oraz } \delta_m \leq \varepsilon_T - \sum_m^T \epsilon_\Delta \tag{10.29}$$

10.3 Modyfikacje poprawiające wydajność uniformizacji dla modeli narodzin i śmierci

Zastosowania modeli CTMC dla systemów kolejkowych proponowane w literaturze są na ogół ograniczone do relatywnie niewielkich systemów – wynika to ze wspomnianej w sekcji 10.1.1 złożoności obliczeniowej rosnącej z kwadratem wielkości systemu. Zagadnienie poprawy wydajności algorytmu uniformizacji dla stosowanych w praktyce modeli CTMC znalazło odzwierciedlenie w wielu publikacjach – przegląd stosowanych w tym celu metod można znaleźć m.in. w (Moorsel and Haverkort, 1996).

Jedną z takich możliwości jest redukcja współczynnika uniformizacji, dzięki identyfikacji stanów, które mogą być osiągnięte w trakcie analizowanego okresu czasu. Na przykład (Van Moorsel and Sanders, 1997) porównuje standardowy algorytm uniformizacji z tzw. uniformizacją adaptywną (ang. *adaptive uniformization* - AU vs. *standard uniformization* - SU), w której dyskretny proces Poissona zostaje zastąpiony procesem narodzin, który generuje dla zadanego odcinka czasu co najwyżej tą samą liczbę zdarzeń, w wyniku czego zostaje zredukowany współczynnik uniformizacji, a co za tym idzie liczba niezbędnych ewolucji podporządkowanego DTMC. Dodatkowo autorzy identyfikują sytuacje, w których koszt obliczania *jump probabilities* procesu narodzin przewyższa oszczędności wynikające ze zredukowanej liczby iteracji w celu obliczania ewolucji DTMC, tzn. zastąpienie SU przez AU nie daje poprawy efektywności.

Inną możliwością, prowadzącą również do zmniejszenia współczynnika uniformizacji, a co za tym idzie mniejszej wartości górnego ograniczenia k, jest redukcja stanów zaproponowana w (Henzinger, Mateescu, and Wolf, 2009) i rozwinięta w (Didier et al., 2009). Obydwie prace poświęcone są modelom CTMC reakcji chemicznych (ang. *chemical master equations*), które charakteryzują się nieskończoną przestrzenią stanów, jak również nieograniczonym zakresem intensywności przejść, i z tego powodu nie mogą być analizowane przy pomocy

standardowych algorytmów numerycznych. Obejściem tego ograniczenia zaproponowanym przez autorów jest analiza ewolucji systemu w krokach, w których zbiór rozważanych stanów jest ograniczony do jedynie takich, których prawdopodobieństwo wystąpienia jest znaczące. Podobne podejście może być również zastosowane dla systemów o skończonej wielkości, w których liczba stanów o znaczącym prawdopodobieństwie wystąpienia jest relatywnie mała w porównaniu z wielkością systemu, w szczególności w interesującym nas przypadku Markowskich kolejkowych modeli narodzin i śmierci. Poprzez redukcję zbioru uwzględnianych stanów do $\hat{K} \ll K$ zmniejsza się również rozmiar wektora prawdopodobieństw stanów ν oraz macierzy prawdopodobieństw przejść P a co za tym idzie nakład pojedynczej operacji mnożenia wektora i macierzy do $O(\hat{K})$.

W przypadku gdy analizowane są modele posiadające jeden stan ustalony (np. kolejkowe modele o ograniczonej pojemności) to dla $\nu_n(i)$, $n = (0, \ldots, K)$ jak w 10.6, oraz

$$\nu_n(\infty) = \lim_{i \to \infty} \nu_n(i) \tag{10.30}$$

będącym rozkładem ustalonym (granicznym) dla $\nu(i)$), wartość dla dowolnego $\nu_n(i), i \leq k$, gdzie k jest górnym ograniczeniem, będzie dla każdego $n = (0, \ldots, K)$ zawsze pomiędzy $\nu_n(\infty)$ and $\pi_n(0)$, gdzie $\pi(0)$ jest rozkładem początkowym dla analizowanego odcinka czasu (kroku). Dodatkowo, można założyć, że masa prawdopodobieństwa wektora stanu ustalonego jest skoncentrowana w pobliżu wartości oczekiwanej, co jest spełnione w przypadku modeli narodzin i śmierci, dla których $\rho_n = \lambda_n/\mu_{n+1} < 1$ (np. modele z wykładniczym procesem abandonment). Pozwala to na redukcję zbioru stanów do $\hat{K} = (s, \ldots, r)$, gdzie $s = min\{s_0, s_\infty\}$ jest pierwszym a $r = max\{r_0, r_\infty\}$ ostatnim stanem z istotnym prawdopodobieństwem wystąpienia, określone dla wybranego ograniczenia błędu ϵ_{sr} takiej redukcji stanów przez:

$$\sum_{n=0}^{s_0-1} \pi_n(0) \leq \frac{\epsilon_{sr}}{2}, \sum_{n=r_0+1}^{K} \pi_n(0) \leq \frac{\epsilon_{sr}}{2} \tag{10.31}$$

$$\sum_{n=0}^{s_\infty-1} \nu_n(\infty) \leq \frac{\epsilon_{sr}}{2}, \sum_{n=r_\infty+1}^{K} \nu_n(\infty) \leq \frac{\epsilon_{sr}}{2} \tag{10.32}$$

Zaproponowana heurystyka nie jest optymalna, szczególnie w przypadku kiedy rozkład prawdopodobieństw stanów na końcu kroku różni się znacząco od stanu ustalonego. Niemniej jednak jej koszt obliczeniowy jest praktycznie pomijalny w przypadku wykorzystania dla zawartej w dysertacji propozycji modyfikacji algorytmu uniformizacji z wykrywaniem stacjonarności, którego wyznaczony z góry stan ustalony można wykorzystać. Ponieważ w większości sytuacji analizowanych m.in. przez (Green, Kolesar, and Soares, 2003) lub (Ingolfsson et al., 2010), przybliżenia stacjonarne dają zadowalające rezultaty, można oczekiwać, że sytuacje, w których będzie możliwe aproksymowanie w granicach dopuszczalnego błędu dokładnego nieustalonego stanu systemu jego stanem ustalonym, będą stanowiły znaczącą część analizowanych przedziałów czasu, a co za tym idzie, proponowane w dysertacji modyfikacje powinny przynosić znaczącą poprawę efektywności w zastosowaniach praktycznych. W szczególności do zastosowań takich jak weryfikacja i dalsza optymalizacja planów obsady uzyskanych za pomocą łatwych obliczeniowo przybliżeń stacjonarnych. Natomiast

w sytuacjach, w których stosowanie przybliżeń stacjonarnych prowadziłoby do uzyskania błędnych wyników (na przykład dla systemów o dużej zmienności, długich czasach obsługi lub dużych rozmiarów) nieco zwiększony w takim wypadku nakład obliczeniowy wydaje się być uzasadniony.

Kolejną możliwością znaczącej poprawy wydajności obliczeniowej 10.6 jest zaproponowana przez Didier et al., 2009 oraz Andreychenko et al., 2010 konstrukcja macierzy P w locie (ang. *on the fly*). Pozwala ona na rezygnację z przechowywania P w pamięci poprzez wyznaczanie niezbędnych do kalkulacji kolejnych ewolucji DTMC prawdopodobieństw przejść bezpośrednio z równań narodzin i śmierci danego modelu. W związku z tym, że nowoczesne CPU potrafią, poprzez wykorzystanie kolejkowania i sprzętowego zrównoleglenia, wykonać znaczącą liczbę operacji arytmetycznych w czasie wymaganym dla pojedyńczej operacji dostępu do pamięci operacyjnej (np. Burger, Goodman, and Kagi, 1996 lub standardowe podręczniki zajmujące się współczesnymi architekturami CPU), zmniejszenie zapotrzebowania algorytmu na pamięć kosztem dodatkowych operacji arytmetycznych może przynieść znaczące przyspieszenie jego pracy.

10.4 Eksperymenty numeryczne

Algorytm zaproponowany w rozdziale trzecim został zaimplementowany w języku C++ a następnie przetestowany przy użyciu przykładów symulujących typowe sytuacje, powodujące zmianę modelu w czasie, występujące w rzeczywistych call center. W szczególności w rozdziale czwartym przetestowany został wpływ skokowej zmiany obsady przy stałym natężeniu przychodzących żądań obsługi oraz ciągłych zmian natężenia procesu przychodzących żądań obsługi, występujących z różną intensywnością dla stałej liczby serwerów. Na przykładzie o stałym natężeniu przychodzących żądań obsługi pokazany został błąd wynikający z aproksymacji stanu nieustalonego na koniec analizowanego okresu czasu (kroku) rozkładem stacjonarnym. Błąd ten zgodnie z oczekiwaniami pozostawał znacząco mniejszy niż jego ograniczenie δ wykorzystywane do podjęcia decyzji o aproksymacji w danym kroku zgodnie z (10.23).

Poprawa efektywności obliczeniowej, będąca rezultatem zastosowania proponowanych w dysertacji modyfikacji algorytmu, wynika bezpośrednio ze zmniejszonej liczby iteracji podporządkowanego wektora DTMC (operacji mnożenia wektor macierz) – w szczególności, jeżeli liczba iteracji potrzebna do podjęcia decyzji o aproksymacji rozwiązania danego kroku wartością stanu ustalonego jest znacząco mniejsza od liczby iteracji k (górnego ograniczenia) potrzebnych do jego wyznaczenia w algorytmie bez wykrywania stacjonarności. Oszczędność ta jest proporcjonalna do długości odcinków czasu w których stan (zmiennego w czasie) systemu jest zbliżony do stanu stacjonarnego. Wyniki eksperymentów przedstawione w dysertacji potwierdzają wnioski zawarte m.in. w (Green, Kolesar, and Svoronos, 1991) i (Green, Kolesar, and Soares, 2001) w których autorzy analizowali dokładność przybliżeń stacjonarnych dla zmiennych w czasie systemów call center. W szczególności poprawa efektywności (wynikająca ze stacjonarnego przybliżenia w granicach z góry założonego błędu) jest największa w systemach o małej zmienności i krótkich czasach obsługi, maleje natomiast wraz z wielkością systemu.

Dodatkowo dla modelu o stałej liczbie serwerów przetestowany został wpływ,

zaproponowanej w rozdziale trzecim, metody redukcji stanów dla złożoności obliczeniowej w funkcji wielkości modelowanego systemu.

W celu oceny możliwej do uzyskania poprawy efektywności dla zastosowań praktycznych, w rozdziale piątym pokazane zostało zastosowanie zaproponowanego w pracy algorytmu dla obliczenia wartości service level dla realistycznej prognozy ruchu wraz z przykładową realistyczną obsadą – analogicznie do pojedynczego kroku metody optymalizacji obsady zaproponowanego przez (Ingolfsson et al., 2010) (przedstawionego w rozdziale drugim). Zestaw danych do stworzenia prognozy ruchu, na podstawie danych rzeczywistego call center średniej wielkości amerykańskiego banku udostępnionych przez (Mandelbaum and Zeltyn, 2013), został wybrany identycznie jak w propozycji w (Kim and Whitt, 2013). Dane przykładowej obsady są identyczne z obsadą z dnia 26 maja 2001 znajdującego się również w ww. zestawie danych, który został wybrany ze względu na zbieżność liczby przychodzących telefonów w tym dniu z wynikami prognozy opartej o dane całego zestawu danych (18-tu dni roboczych w maju 2001). Otrzymany wynik to ok. 50% redukcja czasu obliczeń w porównaniu z algorytmem nie wykorzystującym wykrywania stacjonarności. Dodatkowo, podobnie jak w rozdziale czwartym, eksperyment został powtórzony dla systemów o większym rozmiarze uzyskanych przez 10-cio i 100-krotne przeskalowanie oryginalnych danych obsady i prognozy ruchu, w celu oceny poprawy efektywności wynikającej z zastosowania redukcji stanów.

Rozdział 11

Wnioski

Praca poświęcona jest modelowaniu systemów call center przy użyciu niejednorodnych w czasie łańcuchów Markowa z czasem ciągłym (CTMCs). Modele kolejkowe są powszechnie stosowane w praktyce zarządzania operacyjnego systemów obsługi, w tym w szczególności organizacji świadczących usługi na odległość takich jak call center. Ze względu na dominującą rolę kosztów osobowych, dokładność modeli używanych do celów planowania obsady ma duże znaczenie dla wyników ekonomicznych tego typu organizacji. Modele pozwalające na dokładniejsze odwzorowanie rzeczywistych systemów, w szczególności ich zmienności w czasie oraz innych niż Markowskie rozkładów czasów obsługi i charakterystyk opuszczania systemu (abandonment), niż powszechnie stosowane stacjonarne przybliżenia oparte na modelach Erlang-C i Palm/Erlang-A są przedmiotem zainteresowania wielu ośrodków badawczych. Niestety proponowane dotychczas w literaturze metody zastosowania modeli odwzorowujących stany nieustalone zmiennego w czasie rzeczywistego systemu kolejkowego, jak również dokładniejsze odwzorowanie innych charakterystyk modelowanego systemu były uważane za możliwe do wykorzystania jedynie do modelowania relatywnie małych rzeczywistych systemów – głównie ze względu na ich złożoność obliczeniową.

W pracy zaproponowane zostały modyfikacje algorytmu uniformizacji, będącego standardową metodą wyznaczania stanów nieustalonych CTMCs. W szczególności wykorzystany został fakt, że rzeczywiste systemy call center mogą być przez znaczącą część czasu aproksymowane przy użyciu ich stanów stacjonarnych - będący również, oprócz niskiej złożoności obliczeniowej i łatwości implementacji, przyczyną szerokiego stosowania w praktyce modeli opartych na przybliżeniach stacjonarnych. W pracy zmodyfikowany został zaproponowany przez Muppalę i Trivediego mechanizm wykrywania stacjonarności w algorytmie uniformizacji, tak aby uniknąć problemu przedwczesnego jej wykrycia obecny w dotychczasowych jego implementacjach. W szczególności, zamiast proponowanej w literaturze analizy poszczególnych iteracji podporządkowanego DTMC – które nie dają gwarancji ograniczenia błędu uzyskanego przybliżenia stanu nieustalonego – wykorzystane zostały uzyskane numerycznie właściwości przebiegu funkcji zbieżności, do wyznaczenia z góry dokładnego błędu aproksymacji, wyznaczonym inną metodą dokładnym rozkładem stacjonarnym. Oprócz ograniczenia błędu rozwiązania, propozycja ta pozwala również na znaczącą poprawę efektywności obliczeniowej algorytmu Muppali i Trivediego, dzięki możliwości przerwania iteracji DTMC już po niewielkiej ich liczbie (koniecznej jedynie do określenia przebiegu funkcji zbieżności – bez konieczności obliczania przybliżo-

nego stanu stacjonarnego metodą iteracji DTMC).

Działanie algorytmu zostało przetestowane na przykładach numerycznych, odwzorowujących typowe sytuacje zmienności w czasie rzeczywistych systemów call center, które potwierdziły przewidywane właściwości proponowanego algorytmu. Dodatkowo, dla przykładu ze zmiennym natężeniem ruchu przychodzącego został przetestowany wpływ dodatkowych modyfikacji algorytmu, wykorzystujących specyficzne właściwości kolejkowych modeli narodzin i śmierci, na złożoność obliczeniową algorytmu w funkcji wielkości analizowanego systemu.

Całościowy wpływ zastosowanych zmian na efektywność proponowanego rozwiązania został przetestowany również przy użyciu danych rzeczywistego call center amerykańskiego banku udostępnionych przez prof. Avishai Mandelbauma z Uniwersytetu Technion w Hajfie. Dla call center średniej wielkości (tzn. ze średnią liczbą ok. 200-300 i maksymalną ok. 500 dostępnych agentów), modyfikacje poprawiające efektywność operacji mnożenia wektora i macierzy pozwoliły na ok. 60-krotne zmniejszenie czasu obliczeń w porównaniu z oryginalnym algorytmem uniformizacji. Zastosowanie modyfikacji wykrywania stacjonarności pozwoliło na dodatkowe 2-krotne zwiększenie efektywności obliczeniowej. Osiągnięta redukcja nakładu obliczeniowego jest szczególnie istotna dla zastosowania przy planowaniu obsady (np. z wykorzystaniem programowania całkowitoliczbowego zaproponowanego w Ingolfsson et al., 2010), gdzie konieczne jest wielokrotne wyznaczanie wartości service level przy zmieniającej się obsadzie. Zmniejszenie złożoności obliczeniowej w funkcji wielkości systemu pozwala również na wykorzystanie dla systemów o dowolnej spotykanej w rzeczywistości wielkości. W porównaniu z dotychczas prezentowanymi w literaturze metodami wykorzystującymi niejednorodne w czasie modele CTMC, które były ograniczone jedynie do modelu M/M/S, proponowana metoda umożliwia stosowanie dowolnych Markowskich kolejkowych modeli narodzin i śmierci, co zostało zademonstrowane przy wykorzystaniu realistycznego zależnego od stanu Markowskiego kolejkowego modelu narodzin i śmierci – uwzględniającego blokowanie przychodzących rozmów na skutek ograniczonej pojemności systemu, natychmiastową rezygnację z obsługi po stwierdzeniu faktu konieczności oczekiwania (banking) oraz rezygnację z obsługi po upływie czasu cierpliwości (abandonment), który był dotychczas wykorzystywany jedynie jako przybliżenie stacjonarne.

11.1 Kierunki dalszych badań

W związku z tym, że proponowane w dysertacji modyfikacje metody Muppali i Trivediego nie są ograniczone do kolejkowych modeli systemów obsługi, będących przedmiotem zainteresowania pracy, mogą być one wykorzystane również do analizy niejednorodnych w czasie CTMCs, w przypadku których można oczekiwać, że ich stan przez znaczącą część analizowanego okresu czasu będzie zbliżony do stacjonarnego. Jest to możliwe w szczególności dla modeli, dla których istnieją efektywne metody wyznaczenia ich dokładnych stanów ustalonych z góry (np. przy wykorzystaniu równań równowagi szczegółowej).

Innym interesującym obszarem dalszych badań, wykorzystującym stworzone w niniejszej pracy narzędzia, może być ocena adekwatności bardziej dokładnych modeli rzeczywistych systemów obsługi, proponowanych w dotychczasowej literaturze jedynie jako rozwiązania stacjonarne, również, dla uwzględniających stany nieustalone modeli zmiennych w czasie.

Bibliografia

Aksin, Zeynep, Mor Armony, and Vijay Mehrotra (2007). "The Modern Call Center: A Multi-Disciplinary Perspective on Operations Management Research". In: *Production and Operations Management* 16.6, pp. 665–688. DOI: 10.1111/j.1937-5956.2007.tb00288.x.

Andreychenko, Aleksandr, Pepijn Crouzen, Linar Mikeev, and Verena Wolf (2010). "On-the-fly uniformization of time-inhomogeneous infinite Markov population models". In: *arXiv preprint arXiv:1006.4425*.

Arns, Markus, Peter Buchholz, and Andriy Panchenko (2010). "On the Numerical Analysis of Inhomogeneous Continuous-Time Markov Chains". In: *INFORMS Journal on Computing* 22.3, pp. 416–432. DOI: 10.1287/ijoc.1090.0357.

Bain, Peter and Phil Taylor (2002). "Consolidation,?Cowboys? and the developing employment relationship in British, Dutch and US call centres". In: *Re-Organising Service Work. Ashgate Publishing Limited*, pp. 42–62.

Bolch, Gunter, Stefan Greiner, Hermann de Meer, and Kishor S. Trivedi (2006). *Queueing Networks and Markov Chains*. John Wiley & Sons, Inc. DOI: 10.1002/0471791571.

Burak, Maciej (2014). "Multi-step Uniformization with Steady-State Detection in Nonstationary M/M/s Queuing Systems". In: *arXiv preprint arXiv:1410.0804*.

— (2015a). "Performance analysis of an inbound call center with time varying arrivals". In: *Zeszyty Naukowe Uniwersytetu Szczecińskiego. Service Management* 15, pp. 5–11. DOI: 10.18276/smt.2015.15-01.

Burak, Maciej Rafał (2015b). "Inhomogeneous CTMC Model of a Call Center with Balking and Abandonment". In: *Studia Informatica* 36.2, pp. 23–34. DOI: 10.21936/si2015_v36.n2.712.

— (2016). "Application of an inhomogeneous CTMC model for a telephone call center". In: *Studia Informatica* 37.2, pp. 15–27. DOI: 10.21936/si2016_v37.n2.759.

— (2017). "Computing discrete Poisson probabilities for uniformization algorithm". In: *Studia Informatica* 38.1B, pp. 77–88. DOI: 10.21936/si2017_v38.n1B.795.

Burak, Maciej Rafał and Przemysław Korytkowski (2018). "Inhomogeneous CTMC Models of Birth-and-Death Systems Solved by Uniformization with Steady-State Detection (in review)". In: *INFORMS Journal on Computing*.

Burger, Doug, James R. Goodman, and Alain Kagi (1996). "Memory bandwidth limitations of future microprocessors". In: *ACM SIGARCH Computer Architecture News* 24.2, pp. 78–89. DOI: 10.1145/232974.232983.

Bylina, Jarosław, Beata Bylina, Andrzej Zoła, and Tomasz Skaraczyński (2009). "A Markovian Model of a Call Center with Time Varying Arrival Rate and

Skill Based Routing". In: *Computer Networks*. Springer Science Business Media, pp. 26–33. DOI: 10.1007/978-3-642-02671-3_4.

Cezik, Mehmet Tolga and Pierre L'Ecuyer (2008). "Staffing Multiskill Call Centers via Linear Programming and Simulation". In: *Management Science* 54.2, pp. 310–323. DOI: 10.1287/mnsc.1070.0824.

Creemers, Stefan, Mieke Defraeye, and Inneke Van Nieuwenhuyse (2014). "G-RAND: A phase-type approximation for the nonstationary queue". In: *Performance Evaluation* 80, pp. 102–123. DOI: 10.1016/j.peva.2014.07.025.

Czachórski, Tadeusz, Jean-Michel Fourneau, Tomasz Nycz, and Ferhan Pekergin (2009). "Diffusion Approximation Model of Multiserver Stations with Losses". In: *Electronic Notes in Theoretical Computer Science* 232, pp. 125–143. DOI: 10.1016/j.entcs.2009.02.054.

Defraeye, Mieke and Inneke Van Nieuwenhuyse (2015). "Staffing and scheduling under nonstationary demand for service: A literature review". In: *Omega*. DOI: 10.1016/j.omega.2015.04.002.

Deslauriers, Alexandre, Pierre L'Ecuyer, Juta Pichitlamken, Armann Ingolfsson, and Athanassios N. Avramidis (2007). "Markov chain models of a telephone call center with call blending". In: *Computers & Operations Research* 34.6, pp. 1616–1645. DOI: 10.1016/j.cor.2005.06.019.

Didier, Frederic, Thomas A. Henzinger, Maria Mateescu, and Verena Wolf (2009). "Fast Adaptive Uniformization of the Chemical Master Equation". In: *2009 International Workshop on High Performance Computational Systems Biology*. IEEE. DOI: 10.1109/hibi.2009.23.

Erlang, Agner K (1917). "Solutions of some problems in the theory of probabilities of significance in automatic telephone exchanges. Electroteknikeren, 13: 5–13, 1917". In: *Danish [English transl. in: PO Elec. Eng. J. 10 (1917–18) 189–197.*

Feldman, Zohar, Avishai Mandelbaum, William A. Massey, and Ward Whitt (2008). "Staffing of Time-Varying Queues to Achieve Time-Stable Performance". In: *Management Science* 54.2, pp. 324–338. DOI: 10.1287/mnsc.1070.0821.

Fox, Bennett L. and Peter W. Glynn (1988). "Computing Poisson probabilities". In: *Commun. ACM* 31.4, pp. 440–445. DOI: 10.1145/42404.42409.

Gans, Noah, Ger Koole, and Avishai Mandelbaum (2003). "Telephone Call Centers: Tutorial, Review, and Research Prospects". In: *Manufacturing & Service Operations Management* 5.2, pp. 79–141. DOI: 10.1287/msom.5.2.79.16071.

Grassmann, Winfried K (1978). "Transient solutions in Markovian Queueing Systems". In: *Computers & Operations Research* 5.2, p. 161. DOI: 10.1016/0305-0548(78)90010-2.

Green, Linda and Peter Kolesar (1991). "The Pointwise Stationary Approximation for Queues with Nonstationary Arrivals". In: *Management Science* 37.1, pp. 84–97. DOI: 10.1287/mnsc.37.1.84.

Green, Linda, Peter Kolesar, and Anthony Svoronos (1991). "Some Effects of Nonstationarity on Multiserver Markovian Queueing Systems". In: *Operations Research* 39.3, pp. 502–511. DOI: 10.1287/opre.39.3.502.

Green, Linda V., Peter J. Kolesar, and João Soares (2001). "Improving the Sipp Approach for Staffing Service Systems That Have Cyclic Demands". In: *Operations Research* 49.4, pp. 549–564. DOI: 10.1287/opre.49.4.549.11228.

Green, Linda V, Peter J Kolesar, and Joao Soares (2003). "An improved heuristic for staffing telephone call centers with limited operating hours". In: *Production and Operations Management* 12.1, pp. 46–61. DOI: 10.1111/j.1937-5956.2003.tb00197.x.

Green, Linda V., Peter J. Kolesar, and Ward Whitt (2007). "Coping with Time-Varying Demand When Setting Staffing Requirements for a Service System". In: *Production and Operations Management* 16.1, pp. 13–39. DOI: 10.1111/j.1937-5956.2007.tb00164.x.

Gross, Donald and Douglas R. Miller (1984). "The Randomization Technique as a Modeling Tool and Solution Procedure for Transient Markov Processes". In: *Operations Research* 32.2, pp. 343–361. DOI: 10.1287/opre.32.2.343.

Henken, Kirsten (2007). *Dynamic contact centers with impatient customers and retrials*. VDM Publishing.

Henzinger, Thomas A., Maria Mateescu, and Verena Wolf (2009). "Sliding Window Abstraction for Infinite Markov Chains". In: *Computer Aided Verification*. Springer Berlin Heidelberg, pp. 337–352. DOI: 10.1007/978-3-642-02658-4_27.

Holman, David, Rosemary Batt, and Ursula Holtgrewe (2007). "The global call center report: International perspectives on management and employment". In:

Ingolfsson, Armann and Ling Tang (2012). "Efficient and Reliable Computation of Birth-Death Process Performance Measures". In: *INFORMS Journal on Computing* 24.1, pp. 29–41. DOI: 10.1287/ijoc.1100.0435.

Ingolfsson, Armann, Elvira Akhmetshina, Susan Budge, Yongyue Li, and Xudong Wu (2007). "A Survey and Experimental Comparison of Service-Level-Approximation Methods for Nonstationary M(t)/M/s(t) Queueing Systems with Exhaustive Discipline". In: *INFORMS Journal on Computing* 19.2, pp. 201–214. DOI: 10.1287/ijoc.1050.0157.

Ingolfsson, Armann, Fernanda Campello, Xudong Wu, and Edgar Cabral (2010). "Combining integer programming and the randomization method to schedule employees". In: *European Journal of Operational Research* 202.1, pp. 153–163. DOI: 10.1016/j.ejor.2009.04.026.

Jensen, Arne (1953). "Markoff chains as an aid in the study of Markoff processes". In: *Scandinavian Actuarial Journal* 1953.sup1, pp. 87–91. DOI: 10.1080/03461238.1953.10419459.

JLLResearch (2017). *Contact Centers Outlook - JLL 2017*. URL: http://www.us.jll.com/united-states/en-us/Research/US-Contact-Centers-Outlook-2017-JLL.pdf (visited on 03/20/2018).

Katoen, Joost-Pieter and Ivan S. Zapreev (2006). "Safe On-The-Fly Steady-State Detection for Time-Bounded Reachability". In: *Third International Conference on the Quantitative Evaluation of Systems - (QEST 06)*. IEEE. DOI: 10.1109/qest.2006.47.

Katoen, Joost-Pieter, Ivan S. Zapreev, Ernst Moritz Hahn, Holger Hermanns, and David N. Jansen (2011). "The ins and outs of the probabilistic model checker MRMC". In: *Performance Evaluation* 68.2, pp. 90–104. DOI: 10.1016/j.peva.2010.04.001.

Kim, Song-Hee and Ward Whitt (2013). "Statistical analysis with Little's law". In: *Operations Research* 61.4, pp. 1030–1045. DOI: 10.1287/opre.2013.1193.

Kwiatkowska, Marta, Gethin Norman, and David Parker (2017). "Probabilistic Model Checking: Advances and Applications". In: *Formal System Verification*. Springer International Publishing, pp. 73–121. DOI: 10.1007/978-3-319-57685-5_3.

Malhotra, Manish, Jogesh K. Muppala, and Kishor S. Trivedi (1994). "Stiffness-tolerant methods for transient analysis of stiff Markov chains". In: *Microelectronics Reliability* 34.11, pp. 1825–1841. DOI: 10.1016/0026-2714(94)90137-6.

Mandelbaum, Avishai and Sergey Zeltyn (2007). "Service engineering in action: the Palm/Erlang-A queue, with applications to call centers". In: *Advances in services innovations*. Springer Berlin Heidelberg, pp. 17–45. DOI: 10.1007/978-3-540-29860-1_2.

— (2013). "Data-stories about (im)patient customers in tele-queues". In: *Queueing Systems* 75.2-4, pp. 115–146. DOI: 10.1007/s11134-013-9354-x.

Moorsel, AAD P.A. Van and B.R. Haverkort (1996). "Probabilistic evaluation for the analytical solution of large Markov models: Algorithms and tool support". In: *Microelectronics Reliability* 36.6, pp. 733–755. DOI: 10.1016/0026-2714(95)00193-x.

Muppala, Jogesh K and Kishor S Trivedi (1992). "Numerical transient solution of finite Markovian queueing systems". In: *OXFORD STATISTICAL SCIENCE SERIES*, pp. 262–262.

Reibman, Andrew and Kishor Trivedi (1988). "Numerical transient analysis of Markov models". In: *Computers & Operations Research* 15.1, pp. 19–36. DOI: 10.1016/0305-0548(88)90026-3.

Rindos, Andy, Steven Woolet, Ioannis Viniotis, and Kishor Trivedi (1995). "Exact Methods for the Transient Analysis of Nonhomogeneous Continuous Time Markov Chains". In: *Computations with Markov Chains*. Springer US, pp. 121–133. DOI: 10.1007/978-1-4615-2241-6_8.

Schwarz, Justus Arne, Gregor Selinka, and Raik Stolletz (2016). "Performance analysis of time-dependent queueing systems: Survey and classification". In: *Omega* 63, pp. 170–189. DOI: 10.1016/j.omega.2015.10.013.

Stewart, William J (2009). *Probability, Markov chains, queues, and simulation: the mathematical basis of performance modeling*. Princeton University Press.

Van Moorsel, Aad PA and William H Sanders (1997). "Transient solution of Markov models by combining adaptive and standard uniformization". In: *IEEE Transactions on Reliability* 46.3, pp. 430–440. DOI: 10.1109/24.664016.

Whitt, Ward (2006). "Fluid Models for Multiserver Queues with Abandonments". In: *Operations Research* 54.1, pp. 37–54. DOI: 10.1287/opre.1050.0227.

Younes, Håkan L. S., Marta Kwiatkowska, Gethin Norman, and David Parker (2006). "Numerical vs. statistical probabilistic model checking". In: *International Journal on Software Tools for Technology Transfer* 8.3, pp. 216–228. DOI: 10.1007/s10009-005-0187-8.

Zapreev, I.S. (2008). "Model checking Markov Chains: Techniques and Tools". PhD thesis. DOI: 10.3990/1.9789085702986.